Nucleic Acids Research, 2023, **51**, e110
https://doi.org/10.1093/nar/gkad929
Advance access publication date: 27 October 2023
Methods

A task-specific encoding algorithm for RNAs and RNA-associated interactions based on convolutional autoencoder

Yunxia Wang[1,†], Ziqi Pan[1,†], Minjie Mou[1], Weiqi Xia[1], Hongning Zhang[1], Hanyu Zhang[1], Jin Liu[1], Lingyan Zheng[1,2], Yongchao Luo[1], Hanqi Zheng[1], Xinyuan Yu[1], Xichen Lian[1], Zhenyu Zeng[2], Zhaorong Li[2], Bing Zhang[2], Mingyue Zheng[1,3], Honglin Li[1,4], Tingjun Hou[1] and Feng Zhu[1,2,5,*]

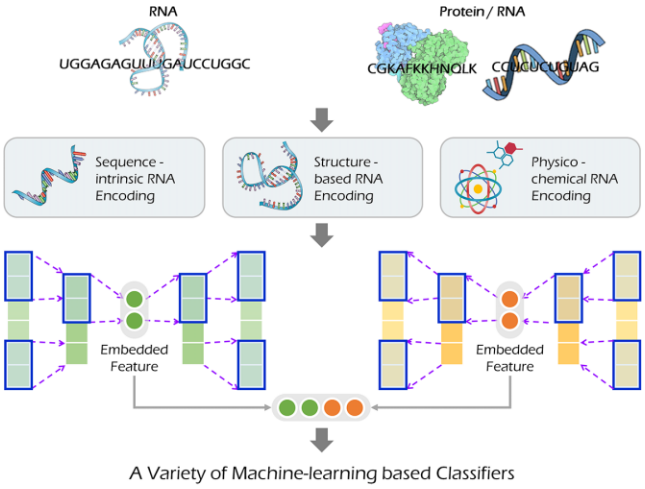
[1]College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Polytechnic Institute, Zhejiang University, Hangzhou 310058, China
[2]Innovation Institute for Artificial Intelligence in Medicine of Zhejiang University, Alibaba-ZJU Joint Research Center of Future Digital Healthcare, Hangzhou 330110, China
[3]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China
[4]School of Pharmacy, East China University of Science and Technology, Shanghai 200237, China
[5]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang, China

[*]To whom correspondence should be addressed. Tel: +86 571 88208444; Fax: +86 571 88208444; Emails: zhufeng@zju.edu.cn; prof.zhufeng@gmail.com
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

RNAs play essential roles in diverse physiological and pathological processes by interacting with other molecules (RNA/protein/compound), and various computational methods are available for identifying these interactions. However, the encoding features provided by existing methods are limited and the existing tools does not offer an effective way to integrate the interacting partners. In this study, a task-specific encoding algorithm for RNAs and RNA-associated interactions was therefore developed. This new algorithm was unique in (*a*) realizing comprehensive RNA feature encoding by introducing a great many of novel features and (*b*) enabling task-specific integration of interacting partners using *convolutional autoencoder*-directed feature embedding. Compared with existing methods/tools, this novel algorithm demonstrated superior performances in diverse benchmark testing studies. This algorithm together with its source code could be readily accessed by all user at: https://idrblab.org/corain/ and https://github.com/idrblab/corain/.

Graphical abstract



Received: May 10, 2022. **Revised:** August 1, 2023. **Editorial Decision:** September 26, 2023. **Accepted:** October 10, 2023

Introduction

RNAs play essential roles in diverse physiological and pathological processes by interacting with other molecules, including RNAs, proteins and compounds (1–4). Particularly, RNAs participate in various post-transcriptional processes by binding to other RNA (5–8); RNA–protein interaction is key to cellular homeostasis, and its perturbations can lead to cellular dysfunction/cancerization (9–12); RNAs bind to compounds (e.g. metabolite) to induce genetic/catalytic function variations and modulate cellular metabolism (13–16). So far, three types of experimental method have been adopted to facilitate the discovery of RNA-associated interactions, which included the '*clustered regularly interspaced short palindromic repeat* (CRISPR)', '*immunoprecipitations*' and '*oligo-captures*' (1,2). Since these methods are characterized by *time-consuming* and *resource-intensive*, the discovery of RNA-associated interactions remains extremely difficult (17–19).

With the advent of '*big data*' era (20–22), various computational approaches have therefore been constructed to identify RNA-associated interactions (23–28). There are two typical steps in these approaches (29–33), which included the *encoding of interacting molecules* into a set of computer-recognizable features (29) and the *integration of encoding features* between interacting partners (30). To realize the above processes, many valuable studies have been conducted to facilitate the encoding of RNA/protein/compound (34–36) and integration of encoding features for interacting partners (26,37). Particularly, a variety of functional tools have been constructed. Some focus on predicting RNA–RNA interactions (RRIs), such as *MD-MLI* (7), *lncIBTP* (38), *PmliPred* (6) and *LncMirNet* (39); some others aim at finding RNA–protein interactions (RPIs), such as *CatRAPID* (10), *PRPI–SC* (40), *PLIPCOM* (41) and *XGBPRH* (42); the remaining tools are designed to find RNA-compound interactions (RCIs), such as *LigandRNA* (15), *dSPRINT* (43) and *RNAmigo* (44). These approaches/tools have emerged to be very popular in various cutting-edge directions, and have attracted broad interests from diverse RNA-associated communities (45–47).

The critical features that were identified to indicate the mechanisms underlying RNA-associated interactions were found to be highly dependent on the studied datasets (48,49). As illustrated in Figure 1, three RRI benchmark datasets (from *Arabidopsis thaliana*, *Glycine max* and *Medicago truncatula*) were statistically analyzed (analytic details are shown in Supplementary Methods), and the identified features from different benchmarks are highly inconsistent. In other words, it is demanded to have a comprehensive coverage of encoding features for each RNA, and the bias towards the features identified from any of the three species will lead to a fail in discovering the features for other species. Furthermore, a successful integration of the encoding features asks for the balanced feature dimensions between different types of interacting partners (50,51). However, the encoding features offered by existing tools are restricted to certain types (39–43), which limits the comprehensiveness of encoding features. Moreover, a simple strategy of direct concatenation is frequently adopted by existing tools (39–43), which does not fully evaluate the balanced feature dimensions. In other words, it is urgently needed to have an RNA encoding method that provides not only comprehensive coverage of encoding features but also a concatenation strategy enabling the integration of the interacting partners of balanced feature dimensions.

In this study, a task-specific encoding algorithm for RNAs and RNA-associated interactions was therefore developed. As described in Figure 2, this proposed algorithm is unique in (a) realizing a comprehensive RNA feature encoding by introducing a large number of novel features and (b) enabling a task-specific integration of interacting partners based on autoencoder-directed feature embedding. To validate the effectiveness of this encoding algorithm, additional case studies were conducted. *First*, the performances of this algorithm for predicting RNA-associated interactions were assessed by well-established benchmarks, and systematically compared with existing tools. *Second*, its ability to decipher the mechanism underlying RNA coding potential (a long-standing problem in modern RNA studies) was also assessed. Based on these analyses, this new algorithm had demonstrated superior ability (comparing with all those existing tools) in not only predicting RNA-associated interactions but also revealing RNA coding potential. The online version of this tool is now readily accessible by all users at: https://idrblab.org/corain/, and all the corresponding source codes can be downloaded from: https://github.com/idrblab/corain/.

Materials and methods

Comprehensive and innovative RNA encoding realized in this study

RNA encoding features published by previous works

A total of 380 RNA descriptors were collected, which had been widely used as encoding features supporting modern RNA study. As illustrated in Figure 3, these 380 encoding features could be grouped into eight feature groups (in white font), which included: sequence-intrinsic features (five groups containing 177 features), physico-chemical features (two groups covering 195 features) and structure-based features (one group including eight features). The brief introduction to and an exemplar application of each feature group were described in both Table 1 and Supplementary Table S1. Since those 380 encoding features were previously published, their corresponding feature groups were named as *Traditional Encoding Feature Groups* (*TraEFGs*, colored in white font in Figure 3), and the detailed description of these *TraEFGs* was provided in the Supplementary Methods. Some feature groups had multiple subgroups, and all the features encoded by the subgroups could then be concatenated to a vector. Taking the feature group '*Open Reading Frame*' as an example, diverse subgroups were covered by this study, which included *basic open reading frame features*, *entropy density profiles on ORF*, *measurement of hexamer on ORF*, and so on.

It is worth noting that two-dimensional encoding features of 3 *TraEFGs* (52–54) as well as spatial structure encoding features of another 3 *TraEFGs* (54–56) were reviewed and organized in Table 1 and Supplementary Table S1. Because of the incompatibility among the features of different dimensions and the lack of spatial structure data for most RNAs, this study focused on discussing RNA descriptors of one dimension, which were comprehensively provided in Figure 3.

RNA encoding features newly proposed by this study

Inspired by the protein encoding strategy proposed by our previous study (35), a total of 297 new RNA descriptors were introduced into this work. As illustrated in Figure 3, these novel encoding features could be grouped into six feature

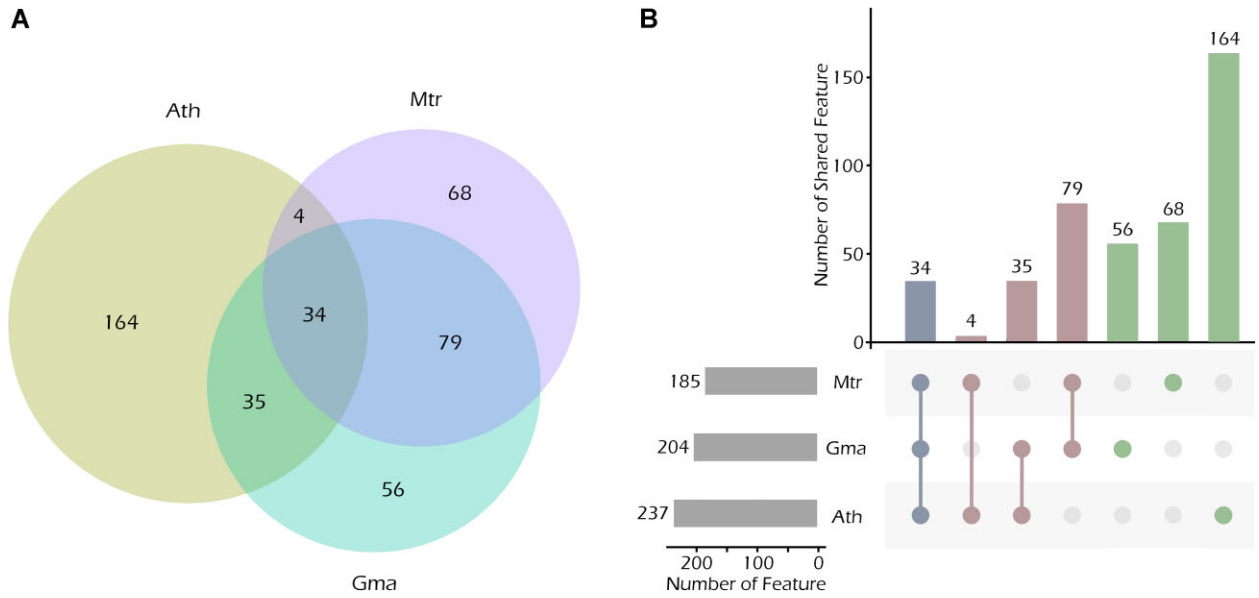


Figure 1. The overlaps of the selected best-performing encoding features of three RNA–RNA interactions (RRIs) datasets. (**A**) the Venn diagram was annotated with the number of exclusive and overlapped features for three RRIs datasets; (**B**) the Upset plot represented the number of features from every overlapping region in (A). Grey strips indicated the number of selected best-performing features of each dataset. Dots and lines indicated the source datasets of the selected features, meaning which ones of the three datasets the selected features belong to. Those histograms indicated the number of selected best-performing features from each overlapping region in (A), with overlapped features of all three datasets illustrated as bar in blue, overlapped features of just two RRIs datasets illustrated as bars in pink and non-overlapped exclusive features illustrated as bars in green.

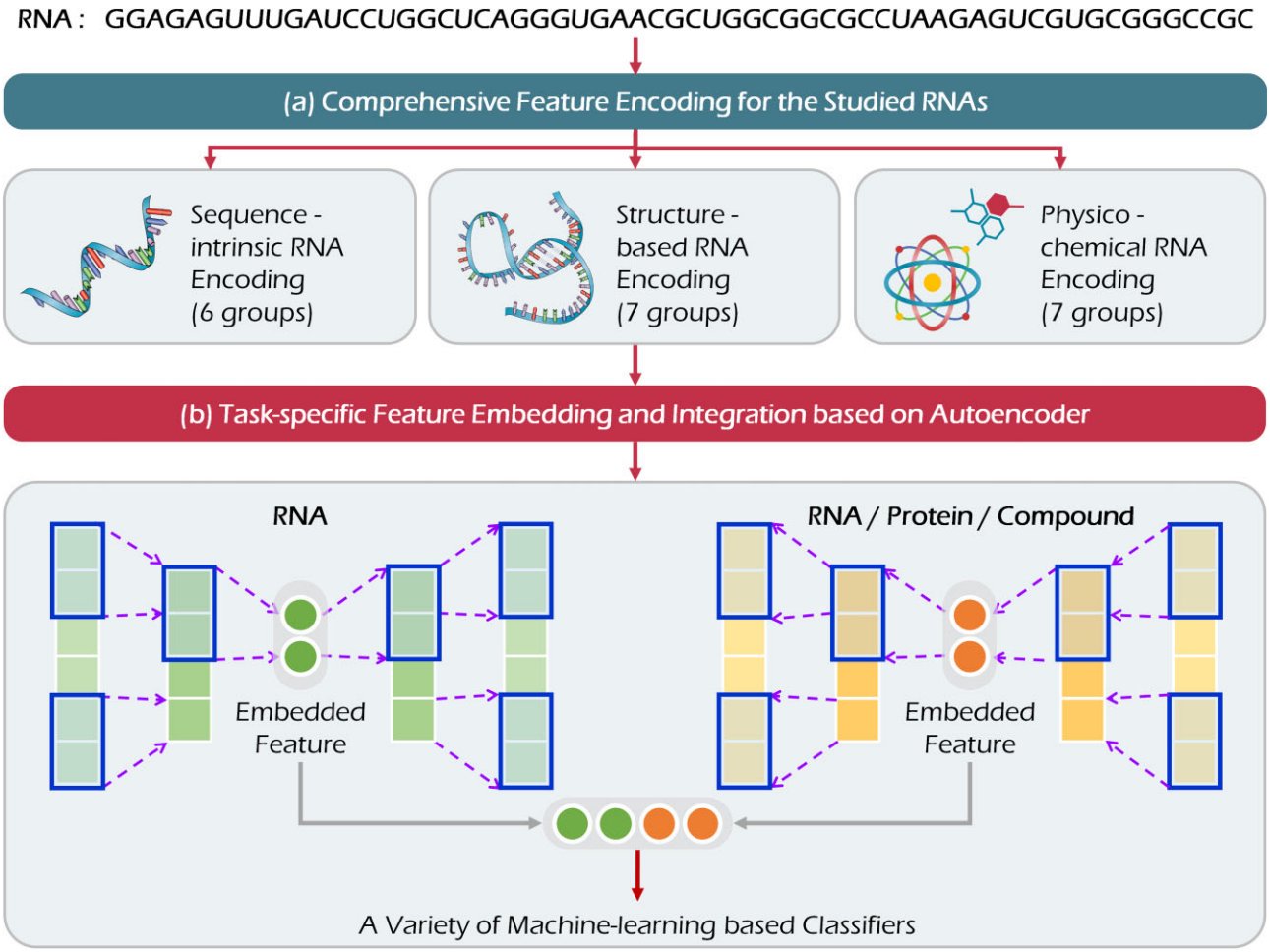


Figure 2. The unique functions provided by CORAIN: (**a**) a comprehensive feature encoding for the studied RNA; (**b**) a task-specific feature embedding and integration based on the autoencoder.

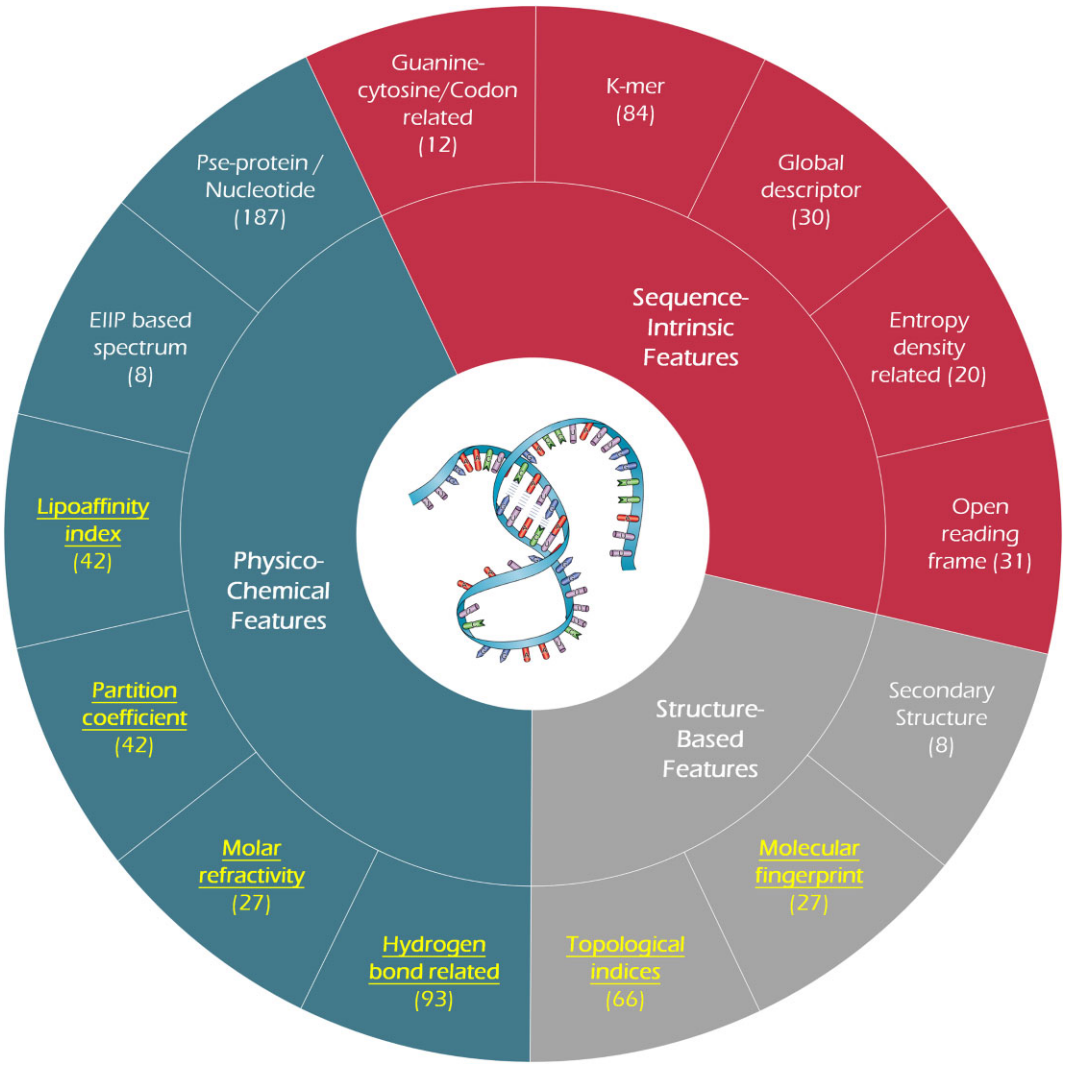


Figure 3. The comprehensive list of RNA encoding feature groups (generating 1D RNA features) applied in this work. The outer-most leaves indicated 14 RNA encoding feature groups, and these groups were divided to three feature classes (sequence-intrinsic, physico-chemical and structure-based that were colored in red, green and grey, respectively). There were two types of feature groups in the outer-most leaves: *New Encoding Feature Groups* (*NewEFGs*, highlighted in yellow bold font) proposed by this study & *Traditional Encoding Feature Groups* (*TraEFGs*, colored in white font) previously published. Numbers in brackets denoted the numbers of features in each feature group. As illustrated, a total of 297 features were from *NewEFG*, and a total of 380 features were from *TraEFG*, which indicated that many *New Encoding Features* from diverse *NewEFGs* were implemented into this study to facilitate encoding of RNAs and RNA-associated Interactions.

groups (in yellow font), which were: physico-chemical features (four groups of 204 features) and structure-based features (two groups of 93 features). A brief introduction to these feature groups was shown in Table 1. Since the 297 encoding features were newly introduced, their corresponding feature groups were entitled to be *New Encoding Feature Groups* (*NewEFGs*, highlighted in yellow bold font in Figure 3), and the detailed description of these *NewEFGs* was also provided in Supplementary Methods. Particularly, based on physico-chemical and structural properties of four nucleotide types that were specified by bases (*adenine*, *guanine*, *cytosine* & *uracil*), the corresponding values were calculated using *PaDEL* (57), which were shown in Supplementary Table S2. Four nucleotide types were classified into two or three groups according to the pre-set threshold values of each physico-chemical and structural property. Then, composition, transition and distribution and k-mer ($k = 1, 2, 3$) were applied to each group of RNAs to calculate their corresponding features. The detailed definition and calculated process for generating these *NewEFGs* were described in detail in Supplementary Methods.

As a result, a total of 677 RNA encoding features were generated by combining all features from eight *TraEFGs* and six *NewEFGs*. To the best of our knowledge, the 677 encoding features were the most comprehensive RNA encoding feature that had ever been provided in the existing tools (25,58,59). Moreover, the encoding features for protein and compound were also made available in our algorithm to facilitate RNA-associated interaction prediction. The detailed information of the way how a protein/compound was encoded was explicitly described in both Supplementary Table S3 and Supplementary Table S4.

Table 1. Description on two typical classes of RNA encoding feature (*Physico-chemical Feature* & *Structure-based Feature*)

Feature group (abbreviation)	Feature type	No. of features (dimensionality)	Brief introduction of 6 *NewEFGs* proposed by this study together with description on the application of 8 *TraEFGs* in previous publications
Feature Class: ***Physico-chemical Features***			
EIIP-based Spectrum (EBS)	*TraEFG*	8 (1D)	Applied to identify lncRNA and predict lncRNA–protein interactions using a classifier generated by various machine learning algorithms (29).
Hydrogen Bond Related (HBR)	*NewEFG*	93 (1D)	A *NewEFG* proposed in this study. The features in this group are calculated based on the properties of hydrogen bond interactions in RNA sequences.
Lipoaffinity Index (SLF)	*NewEFG*	42 (1D)	A *NewEFG* proposed in this study. The features in this group describe the solubility of RNA according to the value of the lipoaffinity index of the bases.
Molar Refractivity (MRA)	*NewEFG*	27 (1D)	A *NewEFG* proposed in this study. The features in this group are based on the molar refractivity to reflect the change of the RNA sequence under polarization.
Partition Coefficient (PCF)	*NewEFG*	42 (1D)	A *NewEFG* proposed in this study. The features in this group characterize the hydrophobic nature of RNA sequences using the value of partition coefficients.
Pse-protein/Nucleotide (PPR)	*TraEFG*	187 (1D)	Applied to predict the noncoding RNAs based on the pseudo protein related features of the RNA sequences using a deep resident network (83).
Sparse Encoding (SEC)	*TraEFG*	1000 × 3 (2D)	Applied to propose a deep learning method for predicting cancer based on generating stacked sparse autoencoders to encode different RNAs (54).
Feature Class: ***Structure-based Features***			
Clash Score (CSC)	*TraEFG*	10 (1D)	Applied to evaluate the prediction performance for RNA 3D structure based on RNA PDB data and structure manipulation (56).
Helical Parameters (HPM)	*TraEFG*	6 (1D)	Applied to identify the RNA-binding sites in a variety of proteins based on the 3D structural information of a large number of RNAs (55).
Molecular Fingerprint (MFP)	*NewEFG*	27 (1D)	A *NewEFG* proposed in this study. The features in this group are calculated based on the molecular fingerprints of four different base structures of RNA.
One-hot Encoding (SCO)	*TraEFG*	1000 × 7 (2D)	Applied to predict the binding site of diverse RNAs or proteins based on deep neural networks with the use of one-hot encoding method (63).
RNA Voxelization (RVL)	*TraEFG*	32 × 32 × 32 (3D)	Applied to systematically assess the prediction performance of RNA tertiary structures using multi-channel convolutional neural network models (54).
Secondary Structure (SST)	*TraEFG*	8 (1D)	Applied to identify lncRNA and predict lncRNA–protein interactions using a classifier generated by various machine learning algorithms (55).
Topological Indices (TGI)	*NewEFG*	66 (1D)	A *NewEFG* proposed in this study. The features in this group are calculated based on the topological relationships among four structures of RNA bases.

There were two types of feature groups: New Encoding Feature Group (NewEFG) proposed by this study & Traditional Encoding Feature Group (TraEFG) published by previous works. For NewEGF, an introduction was provided. For TraEFG, its application in previous study was described. 1D: feature encoded as one-dimensional vector; 2D: feature encoded as two-dimensional matrix; 3D: feature encoded as three-dimensional voxels. Taking all eleven groups of 1D feature as an example, a total of 297 features were from NewEFG, which were significantly larger than that (219) from TraEFG. Such number indicated that a large number of New Encoding Features from diverse NewEFGs were implemented into this study.

A novel strategy proposed for enabling task-specific feature embedding

To eliminate the noise lying in RNA encoding features and modify the feature length for ensuring a balanced integration of the interacting partners, a new self-supervised deep learning framework, *convolutional autoencoder*, was introduced to this study for extracting the informative embedded features. This framework included an encoder module based on the *convolutional neural network* (CNN) and decoder module based on the *deconvolutional algorithm*. As shown in Figure 4, the comprehensive set of RNA encoding features (a total of 677) were *first* obtained and transmitted into the convolutional layer comprising four blocks. Each block contained a convolutional layer followed by a rectified linear unit, a batch normalization layer and a maxpool layer. The changes on the size of feature map was indicated in Figure 4. *Second*, following the convolutional blocks, the feature vectors were flattened and sent into a fully connected layer to customize their lengths and acquire latent feature vectors. *Third*, the latent feature vectors were sent to a fully connected layer and then reshaped for conducting deconvolution. *Fourth*, the deconvolutional blocks would rescale the feature vector into its original size (a total of 677) via the mirror-symmetric paradigm. The changes of the size of feature map was indicated in Figure 4. *Fifth*, a loss calculation based on mean squared error was conducted by which the autoencoder would be retrained to ultimately acquire the final embedded feature. *Sixth*, a similar process was performed to acquire embedded feature for in-

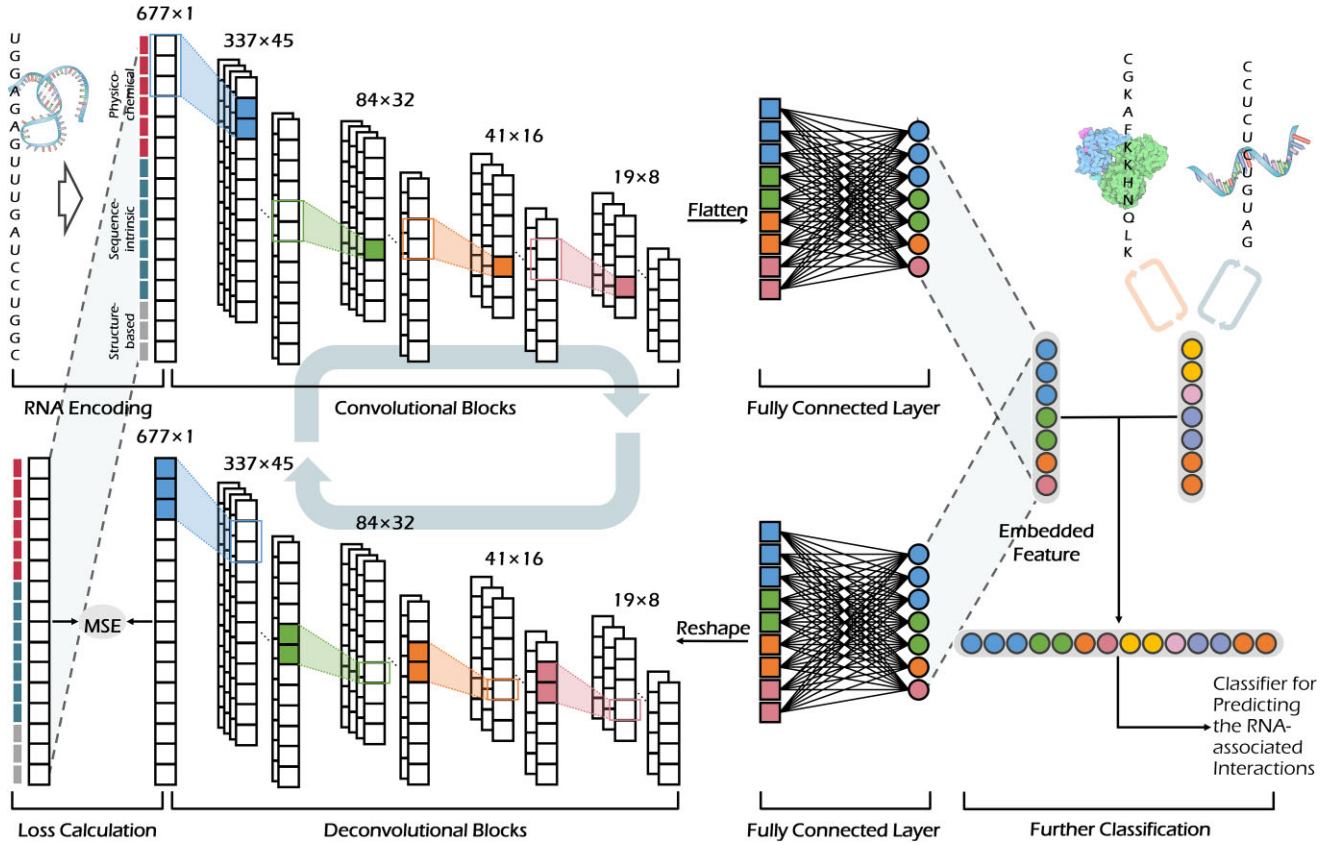


Figure 4. The framework of the self-supervised autoencoder constructed in this study for RNA feature embedding and integration. This autoencoder consisted of an encoder module based on convolutional neural networks and a decoder module based on deconvolutional algorithm. The optimization of autoencoder was achieved by calculating the mean square error (MSE) between the input feature vector and the reconstructed feature vector. The embedded features obtained by the optimized autoencoder would be paired and integrated for predicting RNA-associated interactions. The numbers annotated in the figure denoted the sizes of the feature maps.

teracting partners, which were paired and integrated through concatenation and then used as the input of various classifier to execute downstream prediction tasks.

Assessing encoding performance for RNA and RNA-associated interaction

The encoding algorithm was established by incorporating the pipeline of comprehensive feature encoding, *autoencoder*-based feature embedding and interacting partner integration. In this study, nine benchmark datasets from previous publications were collected, which contained three types of RNA-related prediction tasks: RRIs (6), RPIs (26) and RNA coding potential (25).

A variety of benchmarks for performance assessment

As shown in Table 2, there were three types of benchmark datasets, which were 'benchmarks of RNA–RNA interactions (RRI-bencM)', 'benchmarks of RNA-Protein interactions (RPI-bencM)' and 'benchmarks of RNA coding potential (RCP-bencM)'. Particularly, the RRI-bencM included three datasets of miRNA-lncRNA interactions from three different species: *Arabidopsis thaliana* (Ath), *Glycine max* (Gma) and *Medicago truncatula* (Mtr) (6); the RPI-bencM had five datasets (RPI488, RPI369, RPI1807, RPI2241 & NPInter) from RPITER (26); the RCP-bencM contained one dataset that had 41 917 RefSeq mRNAs (60) and 32 404 Ensembl ncRNAs (61). The detailed information of the studied benchmarks of these three types was provided in Table 2.

Construction of training, validation and test datasets

For RRI-bencM (as shown in Supplementary Figure S1a), cross-species tasks for the prediction of RRIs were conducted on three benchmark datasets from different species (*Ath*, *Gma* and *Mtr*). Specifically, each dataset was divided by 4:1 to training and validation sets. *Then*, the remaining two datasets were used as a test set separately to evaluate the performance of the optimized model. For example, a total of 5000 samples (2500 positive & 2500 negative) from the dataset of *Ath* were divided into the training (2000 positive & 2000 negative) and validation (500 positive & 500 negative) sets. Those remaining two datasets of *Gma* (2500 positive & 2500 negative) and *Mtr* (2500 positive & 2500 negative) were adopted as two independent test sets. As a result, six groups of *Training-Validation-Test* sets were created. For RPI-bencM (Supplementary Figure S1b), the cross-dataset tasks for the prediction of RPIs were conducted on five datasets (RPI488, RPI369, RPI1807, RPI2241 and NPInter). In the task where RPI488 was used as the test set, the remaining datasets (RPI369, RPI1807, RPI2241 & NPInter) were merged and randomly divided by 4:1 into the training and validation sets. Similarly, the datasets of RPI369, RPI1807, RPI2241 and NPInter were successively used as an independent test set with the remaining four as training and validation sets. As a result, five groups of

Table 2. A total of nine benchmark datasets were collected and assessed in this study, which included three RNA–RNA interaction benchmarks, five RNA–protein interaction benchmarks and one RNA coding potential benchmark

Benchmarks of RNA–RNA interactions		**No. of miRNAs**	**No. of LncRNAs**	**No. of molecule pairs**	**Reference**
Arabidopsis thaliana (Ath)	Interacting Pairs	331	2014	2500	*Bioinformatics*.36:2986, 2020 (6)
	Non-interacting Pairs	266	1964	2500	
Glycine max (Gma)	Interacting Pairs	401	1770	2500	*Bioinformatics*. 36:2986, 2020 (6)
	Non-interacting Pairs	542	171	2500	
Medicago truncatula (Mtr)	Interacting Pairs	335	1986	2500	*Bioinformatics*. 36:2986, 2020 (6)
	Non-interacting Pairs	424	2442	2500	
Benchmarks of RNA-Protein Interactions		**No. of RNAs**	**No. of proteins**	**No. of molecule pairs**	**Reference**
NPInter	Interacting Pairs	4636	449	10 412	*NucleicAcids Res*. 42:D104, 2014 (84)
	Non-interacting Pairs	4636	449	10 412	
RPI1807	Interacting Pairs	1072	1801	1807	*NucleicAcids Res*. 43:1370, 2015 (70)
	Non-interacting Pairs	493	1434	1436	
RPI2241	Interacting Pairs	841	2042	2241	*BMCBioinformatics*. 12:489, 2011 (85)
	Non-interacting Pairs	734	2042	2241	
RPI488	Interacting Pairs	24	212	243	*BMCGenomics*. 17:582, 2016 (86)
	Non-interacting Pairs	16	140	245	
RPI369	Interacting Pairs	332	338	369	*BMCBioinformatics*. 12:489, 2011 (85)
	Non-interacting Pairs	223	338	369	
Benchmarks of RNA Coding Potential		**No. of RNAs**			**Reference**
CPPred	Coding RNAs	41917			*NucleicAcidsRes*. 47: e43, 2021 (25)
	Non-coding RNAs	32404			

The numbers of molecules included in each dataset were shown.

Training-Validation-Test sets were generated. For RCP-bencM (as shown in Supplementary Figure S1c), the only dataset was divided by 4:1 into training and validation sets. The same independent test set as used in the original publication (25) was utilized to evaluate the optimized model's performance.

Three types of tasks used for performance assessment

To investigate the prediction performances of the new encoding algorithm, *minimal-redundancy-maximal-relevance* (mRMR) program together with *incremental feature selection* (IFS) method (62) were applied based on the datasets from RPI-bencM (as shown in Table 2). Particularly, (a) the mRMR program was used to rank all 677 features by assessing the significance and relevance of each feature based on both criteria of *minimum redundancy* and *maximum correlation*; (b) the IFS was applied to generate feature subsets through integrating the features incrementally based on the ranking of 677 features to eventually form 677 feature subsets with each subset containing one more feature than the previous one; (c) via thorough scanning of these feature subsets based on random forest (RF) classifier, an optimal feature subset would ultimately be selected.

The prediction task of RNA–RNA interactions (RRIs)

The performances of our algorithm in the prediction of cross-species RRIs were tested based on three benchmarks of RRI-bencM (as provided in Table 2). The strategy for splitting the training-validation dataset followed the same way as described in the section of '*Construction of Training, Validation and Test Datasets*', and six cross-species RRIs tasks were generated (Supplementary Figure S1a). During the encoding process, each RNA was converted into a vector of 677 features, which were then extracted using *autoencoder* to generate the embedded features. Two embedded features of interacting RNA partners were integrated and sent to the same classifier as that of the previous study (6), which was a hybrid model combining CNN with bidirectional gated recurrent unit (CNN-BiGRU). Moreover, two additional classifiers (*random forest* and CNN) were applied for assessment, which had been adopted in previous publications for RRIs prediction (6,38).

The prediction task of RNA–protein interactions (RPIs)

The performances of the newly proposed algorithm in predicting cross-dataset RPIs were *further* tested by five benchmark datasets of RPI-bencM shown in Table 2. The strategy for splitting the training-validation datasets followed the same way as described in the '*Construction of Training, Validation and Test Datasets*' and five cross-dataset RPIs prediction tasks were created (shown in Supplementary Figure S1b). Each interacting RNA was converted to a vector of 677 features following the process described in the section of '*A Novel Strategy Proposed for Enabling Task-specific Feature Embedding*', and the corresponding interacting proteins were transferred into a vector of 438 features using the same strategy as that was described in previous study (26). Both vectors were extracted by *autoencoder* to obtain their embedded features, which were then paired, integrated and sent to the same classifier (an ensemble model of CNN and stacked auto-encoder) as that of the previous study (26). Moreover, four additional classifiers (XGBoost, *support vector machine*, CNN, *random forest*; shown in Supplementary Table S5) were applied for evaluation, which were adopted in previous publications for RPIs prediction (30,63–65).

The prediction task of RNA coding potentials (RCPs)

The performances of the new algorithm in assessing RNA coding potential were compared with ten existing tools based on one benchmark dataset of RCP-bencM (as described in Table 2). The strategy for splitting training-validation dataset followed the same way as described in the section of '*Construction of Training, Validation and Test Datasets*' (provided in Supplementary Figure S1c). During the encoding process, each RNA was converted into a vector of 677 features, which were extracted using *autoencoder* for generating the embedded features and sent to a subsequent classifier (*support vector machine*) to learn and predict the RNA coding potential.

The metrics used for evaluating the constructed models

The classifier of hybrid model CNN-BiGRU and the ensemble model of CNN and stacked auto-encoder that were adopted in this study were the same as that of two previous studies (6,26), and various hyperparameters (*learning rate*, *batch size* & *dropout rate*) were tuned using grid search. *Random forest* was built based on scikit-learn *Python* package (V-0.24.1), and a hyperparameter (*n_estimator*) was tuned by exhaustive searching through manually specified value range in the hyperparameter space. CNN was constructed based on *TensorFlow* (V-2.3.0) and *Keras* libraries (V-2.4.3), *support vector machine* (SVM) was developed using *Python* package *thundersvm* (V-0.3.3) (66), and XGBoost was built by *Python* package *xgboost* (V-1.4.2). The hyperparameters of these three machine learning methods (*learning rate*, *batch size* and *dropout rate* for CNN; *C*, *gamma* and *kernel function* for SVM; *n_round*, *learning rate* and *max depth* for XGBoost) were tuned using grid search method. Detailed information of these studied hyperparameters for RRIs and RPIs were shown in Supplementary Table S6 and Supplementary Table S7, respectively. The optimized hyperparameters for the RCP task were finally set to $C = 10$ and $gamma = 0.001$. The optimized hyperparameters for the *autoencoder* models of all tasks above were displayed in Supplementary Table S8. During performance evaluations, a variety of standard quantification metrics were applied, which included *accuracy* (ACC), *Matthews correlation coefficient* (MCC), and *the area under the receiver operating characteristic curve* (ROC-AUC) (30).

Differentiating features' importance using the algorithm of permutation

It was very essential to investigate the contributing features identified by the new algorithm from the comprehensive encoding features and take a glimpse at the learning result of the *autoencoder*. The importance scores of 677 features were calculated based on the best-performing *autoencoder* model using the permutation algorithm (67), which had been used in previous publications (68). Particularly, the original error was *first* estimated using the original encoded feature matrix as an input. *Second*, for each feature, permuted feature matrix was reformed by permuting features in the original feature matrix, which discarded the influences of features toward the original matrix. *Third*, the permuted errors were estimated using the predictions from the permuted feature matrix, and the permutation feature importance scores were calculated. *Fourth*, higher importance scores reflected greater difference, which indicated the neglection of feature would negatively influence model performances. Thus, features' importance was ranked by descending scores, and the above assessment of feature importance was conducted in all three tasks carried out in this study.

Results and discussion

Application of the novel Task-specific encoding strategy for RRIs prediction

In current cell biology research, it is popular to conduct inexperienced studies upon little-studied species by leveraging high-quality data from popular species (69). In this work, six tasks of cross-species RRIs prediction were established using RRI-bencM shown in Table 2, which considered the genetic difference across species (6). Particularly, the RRIs of one species were used to train model, which was then applied to predict the RRIs of another species. As a result, six cross-species prediction tasks entitled 'training species-test species' (as shown in Figure 5A) were carried out. The performance of the embedded feature was evaluated and compared with the original features proposed in a previous publication (6) using the same classifier as that of the original study.

As illustrated in Figure 5A, the new algorithm using *autoencoder*-based embedded feature (dash line in orange) outperformed the strategy of original study (6) (dash line in grey) in all six cross-species tasks. The improvements in ACCs made by the new algorithm were shown in Figure 5A (over 5% increase colored in blue). Particularly, in the tasks of *Mtr-Gma*, *Mtr-Ath* and *Gma-Ath*, considerable elevations in ACC values were achieved by 17.0%, 11.8% and 24.4%, respectively. Moreover, the embedded features were investigated by additional machine learning model (CNN and *random forest*), which had been commonly applied to predict the RRIs (6,38,39). As a result, our new encoding algorithm achieved a similar level of performance elevations regardless of the applied machine learning models (Supplementary Figure S2). All in all, the proposed algorithm achieved substantial improvements in all tasks, which indicated that it could successfully extract informative RNA features to effectively enhance its capacity in RRIs prediction.

To assess the contributions of the features from the 14 encoding feature groups (8 *TraEFGs* and 6 *NewEFGs*) of the new algorithm, the importance of a total of 677 encoding features (380 from *TraEFGs* and 297 from *NewEFGs*) was evaluated. Based on the permutation algorithm and mean squared error metric (detailed description was provided in the 4th section of Materials and methods), the importance of all features was ranked to reflect their contribution in final RNA representation. All three RRIs datasets above were evaluated, and those features with the positive importance scores were considered to be of great contribution (as shown in Figure 5).

The importance ranking results for three datasets of *Ath*, *Gma* and *Mtr* were separately presented in Figure 5B–C and D. Taking the Figure 5B as an example, the features from *NewEFGs* (bars in red) were ranked higher than those from *TraEFGs* (bars in green). Particularly, in top-100, top-200 and top-300 encoding features, those *NewEFG* features accounted for 33.0%, 44.5% and 46% respectively, highlighting the crucial roles they played in the functioning of the model. Ranking distribution of *NewEFG* (red shading section) & *TraEFG* (green shading section) features were shown in the density distribution chart, more vividly revealing each's contribution. Similar conclusions can be summarized from Figure 5C and D. In sum, *NewEFG* features accounted for 43.5%, 46.3%

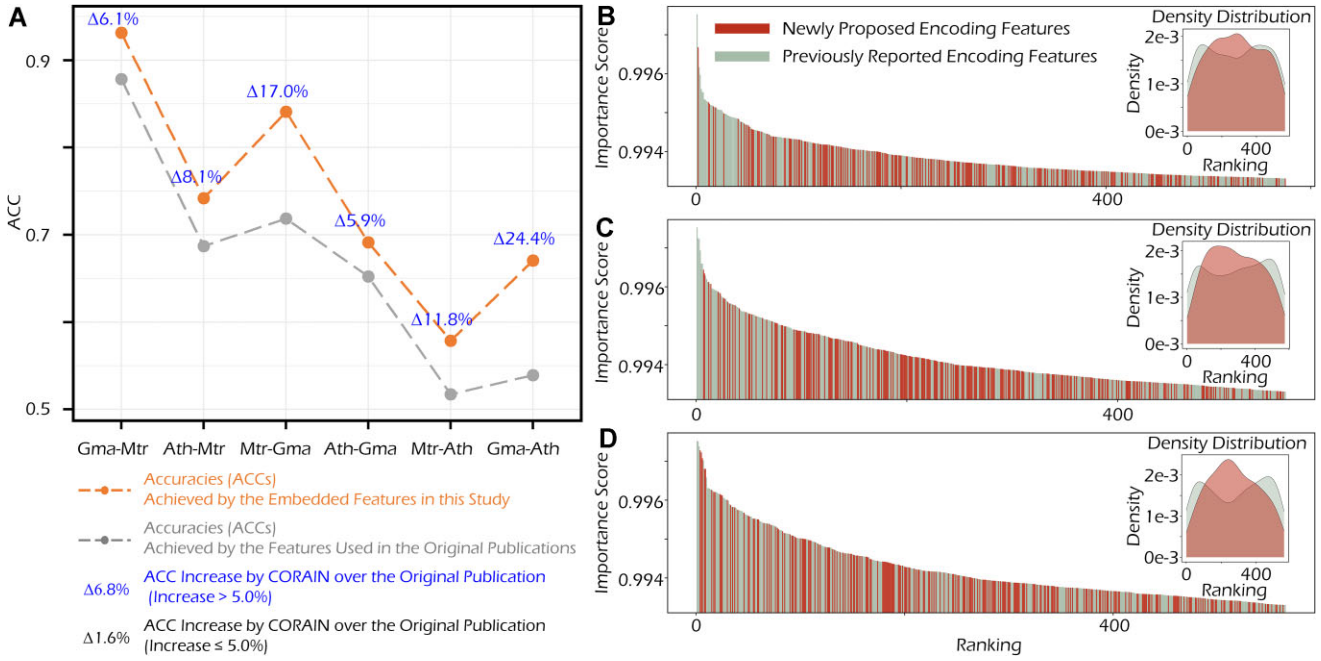


Figure 5. The performance of CORAIN in predicting cross-species RNA–RNA interactions. (**A**) The comparison of performance between CORAIN (dash lines and dots in orange) and the original encoding features from PmliPred (6) (dash lines and dots in grey). The performance was compared using accuracies (ACCs) metric as the indicator and the classifiers from PmliPred (6). The improvements of ACC value made by CORAIN were annotated by Δ (the increases > 5% were colored in blue). RNA–RNA interactions datasets of three species provided in PmliPred (6) were used for performance assessments, which included *Arabidopsis thaliana* (Ath), *Glycine max* (Gma) and *Medicago truncatula* (Mtr). Those six cross-species prediction tasks were entitled 'training species-test species', as provided at the bottom. Taking the 'Mtr-Gma' as an example, the dataset of *Medicago truncatula* (Mtr) was used for training and validation, and the dataset of *Glycine max* (Gma) was adopted for testing. (**B–D**) The importance rankings of contributing RNA features identified by CORAIN. The features in the *New Encoding Feature Groups* (*NewEFGs*) were illustrated as bars in RED, and the features in the *Traditional Encoding Feature Groups* (*TraEFGs*) were indicated as bars in GREEN. These density distribution charts indicated the ranking distribution of *NewEFG* (red shading part) and *TraEFG* features (green shading part). Ranking was conducted for three PmliPred (6) datasets, including (B) dataset of Ath, (C) dataset of Gma and (D) dataset of Mtr.

and 46.2% of all contributing encoding features in the three datasets, which indicated that they were effective complementation to the existing method, improving the predictive performance for cross-species RNA–RNA interactions.

Application of the novel task-specific encoding strategy for RPIs prediction

The prediction of RNA–protein interactions had proved to be a vital scope in the study of cellular process and disease mechanism (70), but remains the challenging research direction. Thus, it was necessity to take a step further and probe the performance of our new algorithm for the prediction of RPIs. To fully illustrate the ability of *autoencoder*-based embedded features, five cross-dataset tasks of RPIs prediction were constructed based on RPI-bencM shown in Table 2, through which limited data could be utilized on much larger scale (26). Particularly, five cross-dataset prediction tasks were generated and investigated using each dataset as the test set, iteratively, with the other four datasets used as a training. Each task was named after its test set, RPI488, RPI369, RPI1807, RPI2241 and NPInter (Figure 6A). The embedded features of our new algorithm were compared with that of previous publication (26) using the same classifier as that of the original study.

As shown in Figure 6A, throughout the five cross-dataset tasks, our new algorithm (dash line in orange) outperformed that of the original publication (26) (dash line in grey). The improvements in ACCs were annotated (over 5% elevation colored in blue). Specifically, in the tasks of RPI369, RPI1807, RPI2241 and NPInter, the considerable enhancement in ACC values was achieved by 8.9%, 11.9%, 10.4% and 23.1%, respectively. Moreover, as described in Supplementary Figure S3, the embedded feature was further investigated using other popular machine learning models (XGBoost, *support vector machine*, CNN, *random forest*; shown in Supplementary Table S5). As a result, our new embedded feature outperformed the original one in all classifiers (shown in Supplementary Figure S3). In sum, the new algorithm made great elevation in all tasks.

To assess the contributions of the features from the 14 encoding feature groups (8 *TraEFGs* and 6 *NewEFGs*) of the new algorithm, the importance of a total of 677 encoding features (380 from *TraEFGs* and 297 from *NewEFGs*) was evaluated. Five RPIs datasets above were evaluated, and those features with the positive importance score were considered to be of great contribution (as shown in Figure 6). The importance ranking results for five datasets (RPI488, RPI369, RPI1807, RPI2241 and NPInter) were presented in Figure 6b–f, respectively. Taking Figure 6b as an example, the *NewEFG* features (bars in red) were ranked higher than the *TraEFG* ones (bars in green). In top-100, top-200 and top-300 encoding features, the *NewEFG* features accounted for 42.0%, 49.0% and 52.3% respectively, highlighting the key roles they played in the functioning of the models. Ranking distribution of *NewEFG* (red shading section) & *TraEFG* (green shading section) features were shown in the density distribution chart,

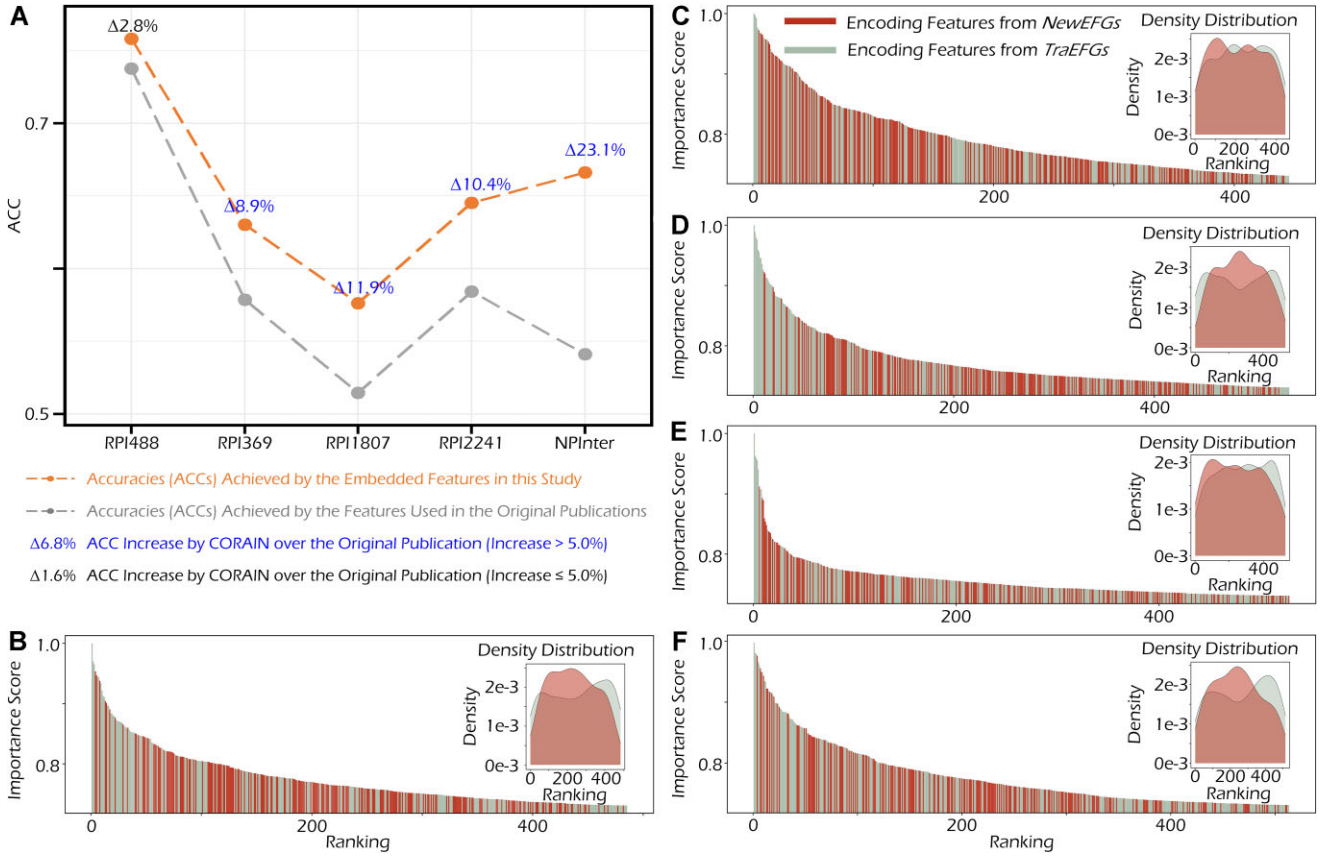


Figure 6. The performance of CORAIN in the prediction of cross-dataset RNA–protein interactions. (**A**) The comparison of performance between CORAIN (dash lines and dots in orange) and the original encoding features of RPITER (26) (dash lines and dots in grey) in the prediction of cross-dataset RNA–protein interactions. The performance was compared using accuracies (ACCs) as the indicator and the classifier from RPITER (26). The improvements of ACC value made by CORAIN were denoted by Δ (the increases > 5% were colored in blue). RNA–protein interactions datasets from the original datasets (RPI488, RPI369, RPI1807, RPI2241 & NPInter) of RPITER (26) were used to generate the training & testing data. These five cross-dataset tasks were named after each's test set, as provided at the bottom. Taking the 'RPI488' as an example, RPI488 was used as testing, and the remaining four datasets were generated as the training and validation sets. (**B–F**) The importance rankings of contributing RNA features identified by CORAIN. The features in *New Encoding Feature Groups* (*NewEFGs*) were illustrated as bars in RED, and the features in the *Traditional Encoding Feature Groups* (*TraEFGs*) were indicated as bars in GREEN. These density distribution charts indicated the ranking distribution of *NewEFG* (red shading part) and *TraEFG* features (green shading part). Ranking was conducted to five distinct RPITER (26) datasets including: (B) RPI488, (C) RPI369, (D) RPI1807, (E) RPI2241 and (F) NPInter.

more vividly revealing each's contribution. Similar conclusions can be summarized from Figure 6C–F. In sum, *NewEFG* features accounted for 47.8%, 50.6%, 48.3%, 42.8% and 45.0% of all encoding features in five datasets, which indicated that they were effective in improving the predictive performance for cross-dataset RNA–protein interactions.

To explain the contribution of features in RNA-related studies, two exemplar datasets ('RPI2241' focusing mostly on the RPIs between ribosomal RNA and protein & 'NPInter' primarily on the RPIs between non-ribosomal RNA and protein) were analyzed. Based on those analyses realized by our new algorithm, the importance scores of all features were calculated for both datasets. As shown in Supplementary Figure S4, those unique features identified from each dataset and their overlap were statistically assessed (those top 10% of all features were evaluated as suggested by previous publication (71). As a result, a variety of features related to '*lipoaffinity*' were identified from RPI2241 dataset, while not being discovered from NPInter. These newly identified features belonged to the *NewEFG* features that were introduced, for the first time, to RNA-related studies by this work, which was named as the *Lipoaffinity Index* feature group. The previous experiments have discovered that the ribosomal RNAs (mostly included in RPI2241) interacted with the lipids, thereby influencing their interactions with proteins (72), which were consistent to the discovery of this work (*Lipoaffinity Index* features). When it came to those overlap features simultaneously identified from both datasets, the feature group of *Hydrogen Bond-related* was discovered in this study, which also belonged to the *NewEFG* feature groups that were introduced, for the first time, to RNA-related studies by this work. Previous studies have proved the crucial roles of hydrogen bonds in multiple RPI complexes (73), which were also highly consistent to the discovery of this work (*Hydrogen Bond-related* features). All in all, our new algorithm could also been applied to uncover the key RNA features underlying the mechanisms of various RNA-related studies.

Application of the encoding algorithm for RNA coding potential prediction

In this study, the ability of the newly proposed algorithm to reveal RNA coding potential (a long-standing problem in modern RNA studies) were further assessed. Particularly, the

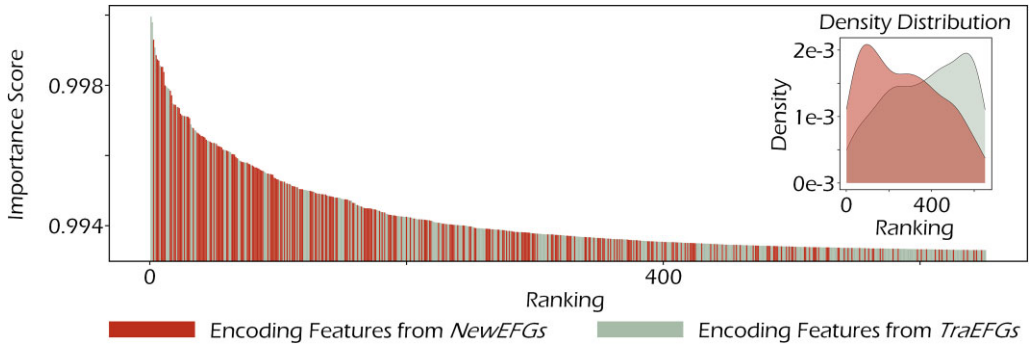


Figure 7. The importance rankings of encoding features identified using CORAIN for predicting RNA coding potential. Those features in *New Encoding Feature Groups* (*NewEFGs*) were shown as bars in RED, and those features in the *Traditional Encoding Feature Groups* (*TraEFGs*) were indicated as bars in GREEN. The density distribution charts indicated the ranking distribution of *NewEFG* (red shading part) and *TraEFG* features (green shading part). Rankings were conducted to the published CPPred (25) dataset.

Table 3. Comparing the performance of CORAIN with that of ten existing SOTA methods for RNA coding potential prediction based on a previously published benchmark dataset (25)

	ACC	MCC	AUC
CPAT	0.943	0.884	0.983
CPC2	0.931	0.863	0.981
CPE-SLDI	0.822	0.680	0.957
CPPred	0.947	0.894	0.987
DeepCPP	0.886	0.774	0.926
FEELnc	0.955	0.911	0.988
mRNN	0.824	0.673	0.940
PLEK	0.966	0.932	0.992
RNAmining	0.899	0.797	0.965
RNAsamba	0.956	0.917	0.992
CORAIN (this study)	**0.971**	**0.941**	**0.995**

These existing methods included CPAT (75), CPC2 (76), CPE-SLDI (77), CPPred (25), DeepCPP (79), FEELnc (80), mRNN (81), PLEK (74), RNAmining (78) and RNAsamba (82). ACC: accuracy; MCC: Matthews correlation coefficient; AUC: area under ROC curve. The highest performance values under three metrics (ACC, MCC and AUC) are highlighted in BOLD font, respectively.

performances of the new algorithm were compared with all existing tools (4,25,74–78) based on the datasets of RCPbencM in Table 2. As shown in Table 3, our new algorithm achieved a significantly better performance comparing with the existing tools: CPAT (75), CPC2 (76), CPE-SLDI (77), CPPred (25), DeepCPP (79), FEELnc (80), mRNN (81), PLEK (74), RNAmining (78), RNAsamba (82), etc. Detailed information of their adopted encoding strategy and applied models was provided in Supplementary Table S9. Particularly, MCCs achieved by new algorithm made were increased by 5.7%, 7.8%, 26.1%, 4.7%, 16.7%, 3.3%, 39.8%, 0.9%, 14.4% and 2.6%, respectively, which showed its ability in revealing the mechanisms underlying the RNA coding potential.

Moreover, it was of great interests to further evaluate what was gained by feature-based encoding methods over non-feature-based ones (especially the mRNN (81)). The mRNN relied heavily on a typical non-feature-based strategy (the one-hot), and its performance on revealing RNA coding potential was described in Table 3. As shown, the feature-based encoding algorithms (including our newly proposed one) were generally better-performed comparing with the non-feature-based one. Such elevation in performance may indicate the improved ability of feature-based encoding in some RNA-related studies. Moreover, due to the explicit description on RNA features, feature-based encoding approaches would have better capacity in explaining the mechanisms underlying an RNA study (better interpretability), when comparing with the non-feature-based ones.

To assess the contributions of the features from the 14 encoding feature groups (8 *TraEFGs* and 6 *NewEFGs*) of the new algorithm, the importance of a total of 677 encoding features (380 from *TraEFGs* and 297 from *NewEFGs*) was evaluated. An RCP benchmark was evaluated, and those features with the positive importance score were considered to be of great contribution (as shown in Figure 7). The importance ranking results for the studied dataset were presented, the *NewEFG* features (bars in red) were ranked higher than the *TraEFG* ones (bars in green). In those top-50, top-100, top-200 and top-300 encoding features, *NewEFG* features accounted for 64.0%, 68.0%, 62.0% and 58.0% respectively, highlighting the key roles they played in the constructed models. Ranking distribution of *NewEFG* (red shading) & *TraEFG* (green shading) features were shown in density distribution chart. All in all, the newly introduced features contributed significantly to the revealing of the mechanism underlying RNA coding potential.

Deployment of an online tool facilitating Task-specific RNA Encoding

To make it usable and accessible to boarder users, an online tool (https://idrblab.org/corain/) was constructed based on our newly proposed strategy. This tool was deployed using the *Python* web framework of *Tornado* (an asynchronous networking library) on a *Linux* webserver implemented with an eight-core CPU of 3.10 GHz and a memory of 64 GB. To the best of our knowledge, this tool was unique in (a) providing the comprehensive set of features for encoding RNAs which is accompanied by additional encoding features for RNA-interacting molecules and (b) realizing a task-specific feature embedding and integration based on autoencoder. The online version of this tool is now readily accessible by all users at: https://idrblab.org/corain/, and all the corresponding source codes can be downloaded from: https://github.com/idrblab/corain/. Additional dataset and software package were required for running these GitHub source codes, which included a dataset file named '*swissprot*' and a software package named '*ncbi-blast-2.9.0+*'. The dataset '*swissprot*' was made downloadable from a separate site (http://idrblab.org/corain/download/swissprot.zip), and the software pack-

age '*ncbi-blast-2.9.0+*' should be downloaded from the website of US NIH (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.9.0/ncbi-blast-2.9.0+-x64-linux.tar.gz).

Conclusions

In this study, a task-specific encoding algorithm for RNAs and RNA-associated interactions was developed. This algorithm was unique in (a) realizing comprehensive RNA feature encoding by introducing numerous novel features, and (b) enabling the task-specific integration of interacting partners using *autoencoder*-directed feature embedding. This novel algorithm had demonstrated superior ability (compared with existing tools) in not only predicting RNA-associated interaction but also revealing RNA coding potential. Our algorithm and its source codes are now accessible by all user at: https://idrblab.org/corain/ and https://github.com/idrblab/corain/.

Data availability

The data underlying this article are available in Zenodo at https://doi.org/10.5281/zenodo.8404705.

Supplementary data

Supplementary Data are available at NAR Online.

Funding

National Natural Science Foundation of China [82373790, 22220102001, U1909208, 81872798]; Natural Science Foundation of Zhejiang Province [LR21H300001]; Leading Talent of the 'Ten Thousand Plan' - National High-Level Talents Special Supports Plan of China; National Key R&D Program of China [2022YFC3400501]; Key R&D Program of Zhejiang Province [2020C03010]; 'Double Top-Class' Universities Projects [181201*194232101]; Fundamental Research Funds for Central University [2018QNA7023]; Alibaba-Zhejiang University Joint Research Center Future Digital Healthcare; Westlake Laboratory (Westlake Laboratory of Life Science & Biomedicine); Alibaba Cloud; Information Technology Center of Zhejiang University. Funding for open access charge: Natural Science Foundation of Zhejiang Province [LR21H300001].

Conflict of interest statement

None declared.

References

1. Ramanathan,M., Porter,D.F. and Khavari,P.A. (2019) Methods to study RNA–protein interactions. *Nat. Methods*, **16**, 225–234.
2. Zhang,Z., Sun,W., Shi,T., Lu,P., Zhuang,M. and Liu,J.L. (2020) Capturing RNA–protein interaction via CRUIS. *Nucleic Acids Res.*, **48**, e52.
3. Gainza,P., Sverrisson,F., Monti,F., Rodola,E., Boscaini,D., Bronstein,M.M. and Correia,B.E. (2020) Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods*, **17**, 184–192.
4. Zhang,S., Amahong,K., Sun,X., Lian,X., Liu,J., Sun,H., Lou,Y., Zhu,F. and Qiu,Y. (2021) The miRNA: a small but powerful RNA for COVID-19. *Brief Bioinform*, **22**, 1137–1149.
5. Van Treeck,B. and Parker,R. (2018) Emerging roles for intermolecular RNA–RNA interactions in RNP assemblies. *Cell*, **174**, 791–802.
6. Kang,Q., Meng,J., Cui,J., Luan,Y. and Chen,M. (2020) PmliPred: a method based on hybrid model and fuzzy decision for plant miRNA-lncRNA interaction prediction. *Bioinformatics*, **36**, 2986–2992.
7. Song,J., Tian,S., Yu,L., Yang,Q., Xing,Y., Zhang,C., Dai,Q. and Duan,X. (2020) MD-MLI: prediction of miRNA-lncRNA interaction by using multiple features and hierarchical deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinform*, **19**, 1724–1733.
8. Zhang,S., Amahong,K., Zhang,C., Li,F., Gao,J., Qiu,Y. and Zhu,F. (2022) RNA–RNA interactions between SARS-CoV-2 and host benefit viral development and evolution during COVID-19 infection. *Brief. Bioinform*, **23**, bbab397.
9. Weidmann,C.A., Mustoe,A.M., Jariwala,P.B., Calabrese,J.M. and Weeks,K.M. (2021) Analysis of RNA–protein networks with RNP-MaP defines functional hubs on RNA. *Nat. Biotechnol.*, **39**, 347–356.
10. Bellucci,M., Agostini,F., Masin,M. and Tartaglia,G.G. (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.
11. Lanjanian,H., Nematzadeh,S., Hosseini,S., Torkamanian-Afshar,M., Kiani,F., Moazzam-Jazi,M., Aydin,N. and Masoudi-Nejad,A. (2021) High-throughput analysis of the interactions between viral proteins and host cell RNAs. *Comput. Biol. Med.*, **135**, 104611.
12. Duan,N., Arroyo,M., Deng,W., Cardoso,M.C. and Leonhardt,H. (2021) Visualization and characterization of RNA–protein interactions in living cells. *Nucleic Acids Res.*, **49**, e107.
13. Meyer,S.M., Williams,C.C., Akahori,Y., Tanaka,T., Aikawa,H., Tong,Y., Childs-Disney,J.L. and Disney,M.D. (2020) Small molecule recognition of disease-relevant RNA structures. *Chem. Soc. Rev.*, **49**, 7167–7199.
14. Warner,K.D., Hajdin,C.E. and Weeks,K.M. (2018) Principles for targeting RNA with drug-like small molecules. *Nat. Rev. Drug Discov.*, **17**, 547–558.
15. Philips,A., Milanowska,K., Lach,G. and Bujnicki,J.M. (2013) LigandRNA: computational predictor of RNA-ligand interactions. *RNA*, **19**, 1605–1616.
16. Sheridan,C. (2021) First small-molecule drug targeting RNA gains momentum. *Nat. Biotechnol.*, **39**, 6–8.
17. Mahmud,S.M.H., Chen,W., Liu,Y., Awal,M.A., Ahmed,K., Rahman,M.H. and Moni,M.A. (2021) PreDTIs: prediction of drug-target interactions based on multiple feature information using gradient boosting framework with data balancing and feature selection techniques. *Brief Bioinform*, **22**, bbab046.
18. Hwang,B., Lee,J.H. and Bang,D. (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp. Mol. Med.*, **50**, 1–14.
19. Wekesa,J.S., Meng,J. and Luan,Y. (2020) A deep learning model for plant lncRNA–protein interaction prediction with graph attention. *Mol. Genet. Genomics*, **295**, 1091–1102.
20. Barquist,L. and Vogel,J. (2015) Accelerating discovery and functional analysis of small RNAs with new technologies. *Annu. Rev. Genet.*, **49**, 367–394.
21. Petti,M., Verrienti,A., Paci,P. and Farina,L. (2021) SEaCorAl: identifying and contrasting the regulation-correlation bias in RNA-seq paired expression data of patient groups. *Comput. Biol. Med.*, **135**, 104567.
22. Jiang,L., Liu,Z., Wang,Z., Su,Y., Wang,Y., Wei,Y., Jiang,Y., Jia,Z., Ma,C., Gang,F., *et al.* (2021) Development of methods for detecting the fate of mesenchymal stem cells regulated by bone bioactive materials. *Bioact Mater*, **6**, 613–626.
23. Roy,S., Sharma,B., Mazid,M.I., Akhand,R.N., Das,M., Marufatuzzahan,M., Chowdhury,T.A., Azim,K.F. and Hasan,M. (2021) Identification and host response interaction study of

SARS-CoV-2 encoded miRNA-like sequences: an in silico approach. *Comput. Biol. Med.*, **134**, 104451.
24. Edera,A.A., Small,I., Milone,D.H. and Sanchez-Puerta,M.V. (2021) Deepred-Mt: deep representation learning for predicting C-to-U RNA editing in plant mitochondria. *Comput. Biol. Med.*, **136**, 104682.
25. Tong,X. and Liu,S. (2019) CPPred: coding potential prediction based on the global description of RNA sequence. *Nucleic Acids Res.*, **47**, e43.
26. Peng,C., Han,S., Zhang,H. and Li,Y. (2019) RPITER: a hierarchical deep learning framework for ncRNA(-)protein interaction prediction. *Int. J. Mol. Sci.*, **20**, 1070.
27. Chauhan,A., Avti,P., Shekhar,N., Prajapat,M., Sarma,P., Bhattacharyya,A., Kumar,S., Kaur,H., Prakash,A. and Medhi,B. (2021) Structural and conformational analysis of SARS CoV 2 N-CTD revealing monomeric and dimeric active sites during the RNA-binding and stabilization: insights towards potential inhibitors for N-CTD. *Comput. Biol. Med.*, **134**, 104495.
28. Chen,K., Xu,H., Lei,Y., Lio,P., Li,Y., Guo,H. and Ali Moni,M. (2021) Integration and interplay of machine learning and bioinformatics approach to identify genetic interaction related to ovarian cancer chemoresistance. *Brief. Bioinform.*, **22**, bbab100.
29. Han,S., Liang,Y., Ma,Q., Xu,Y., Zhang,Y., Du,W., Wang,C. and Li,Y. (2019) LncFinder: an integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property. *Brief. Bioinform.*, **20**, 2009–2027.
30. Hu,H., Zhang,L., Ai,H., Zhang,H., Fan,Y., Zhao,Q. and Liu,H. (2018) HLPI-ensemble: prediction of human lncRNA–protein interactions based on ensemble strategy. *RNA Biol.*, **15**, 797–806.
31. Wang,Q., Wei,L., Guan,X., Wu,Y., Zou,Q. and Ji,Z. (2014) Briefing in family characteristics of microRNAs and their applications in cancer research. *Biochim. Biophys. Acta*, **1844**, 191–197.
32. Zhang,P., Tao,L., Zeng,X., Qin,C., Chen,S., Zhu,F., Li,Z., Jiang,Y., Chen,W. and Chen,Y.Z. (2017) A protein network descriptor server and its use in studying protein, disease, metabolic and drug targeted networks. *Brief. Bioinform.*, **18**, 1057–1070.
33. Zou,Q., Mao,Y., Hu,L., Wu,Y. and Ji,Z. (2014) miRClassify: an advanced web server for miRNA family classification and annotation. *Comput. Biol. Med.*, **45**, 157–160.
34. Nair,A.S. and Sreenadhan,S.P. (2006) A coding measure scheme employing electron-ion interaction pseudopotential (EIIP). *Bioinformation*, **1**, 197–202.
35. Rao,H.B., Zhu,F., Yang,G.B., Li,Z.R. and Chen,Y.Z. (2011) Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.*, **39**, W385–W390.
36. Chen,Z., Zhao,P., Li,F., Marquez-Lago,T.T., Leier,A., Revote,J., Zhu,Y., Powell,D.R., Akutsu,T., Webb,G.I., *et al.* (2020) iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinform.*, **21**, 1047–1057.
37. Tsubaki,M., Tomii,K. and Sese,J. (2019) Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, **35**, 309–318.
38. Zhang,Y., Jia,C. and Kwoh,C.K. (2021) Predicting the interaction biomolecule types for lncRNA: an ensemble deep learning approach. *Brief. Bioinform.*, **22**, bbaa228.
39. Yang,S., Wang,Y., Lin,Y., Shao,D., He,K. and Huang,L. (2020) LncMirNet: predicting lncRNA-miRNA interaction based on deep learning of ribonucleic acid sequences. *Molecules*, **25**, 4372.
40. Zhou,H., Wekesa,J.S., Luan,Y. and Meng,J. (2021) PRPI-SC: an ensemble deep learning model for predicting plant lncRNA–protein interactions. *BMC Bioinf.*, **22**, 415.
41. Deng,L., Wang,J., Xiao,Y., Wang,Z. and Liu,H. (2018) Accurate prediction of protein-lncRNA interactions by diffusion and HeteSim features across heterogeneous network. *BMC Bioinf.*, **19**, 370.
42. Deng,L., Sui,Y. and Zhang,J. (2019) XGBPRH: prediction of binding hot spots at protein(-)RNA interfaces utilizing extreme gradient boosting. *Genes*, **10**, 242.
43. Etzion-Fuchs,A., Todd,D.A. and Singh,M. (2021) dSPRINT: predicting DNA, RNA, ion, peptide and small molecule interaction sites within protein domains. *Nucleic Acids Res.*, **49**, e78.
44. Oliver,C., Mallet,V., Gendron,R.S., Reinharz,V., Hamilton,W.L., Moitessier,N. and Waldispuhl,J. (2020) Augmented base pairing networks encode RNA-small molecule binding preferences. *Nucleic Acids Res.*, **48**, 7690–7699.
45. Song,Z., Lin,J., Su,R., Ji,Y., Jia,R., Li,S., Shan,G. and Huang,C. (2022) EIF3J inhibits translation of a subset of circular RNAs in eukaryotic cells. *Nucleic Acids Res.*, **50**, 11529–11549.
46. Sun,D., Li,Y., Ma,Z., Yan,X., Li,N., Shang,B., Hu,X., Cui,K., Koiwa,H. and Zhang,X. (2021) The epigenetic factor FVE orchestrates cytoplasmic SGS3-DRB4-DCL4 activities to promote transgene silencing in arabidopsis. *Sci. Adv.*, **7**, eabf3898.
47. Corley,M., Burns,M.C. and Yeo,G.W. (2020) How RNA-binding proteins interact with RNA: molecules and mechanisms. *Mol. Cell*, **78**, 9–29.
48. Camperi,J., Moshref,M., Dai,L. and Lee,H.Y. (2022) Physicochemical and functional characterization of differential CRISPR-Cas9 ribonucleoprotein complexes. *Anal. Chem.*, **94**, 1432–1440.
49. Sanchez de Groot,N., Armaos,A., Grana-Montes,R., Alriquet,M., Calloni,G., Vabulas,R.M. and Tartaglia,G.G. (2019) RNA structure drives interaction with proteins. *Nat. Commun.*, **10**, 3246.
50. Wan,X., Wu,X., Wang,D., Tan,X., Liu,X., Fu,Z., Jiang,H., Zheng,M. and Li,X. (2022) An inductive graph neural network model for compound-protein interaction prediction based on a homogeneous graph. *Brief Bioinform*, **23**, bbac073.
51. Zhang,W., Yue,X., Tang,G., Wu,W., Huang,F. and Zhang,X. (2018) SFPEL-LPI: sequence-based feature projection ensemble learning for predicting LncRNA–protein interactions. *PLoS Comput. Biol.*, **14**, e1006616.
52. Fan,X.N., Zhang,S.W., Zhang,S.Y. and Ni,J.J. (2020) LncRNA_Mdeep: an alignment-free predictor for distinguishing long non-coding RNAs from protein-coding transcripts by multimodal deep learning. *Int. J. Mol. Sci.*, **21**, 5222.
53. Zhao,X., Zhang,Y. and Du,X. (2022) DFpin: deep learning-based protein-binding site prediction with feature-based non-redundancy from RNA level. *Comput. Biol. Med.*, **142**, 105216.
54. Xiao,Y., Wu,J., Lin,Z. and Zhao,X. (2018) A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data. *Comput. Methods Programs Biomed.*, **166**, 99–105.
55. Zhang,S., Zhou,J., Hu,H., Gong,H., Chen,L., Cheng,C. and Zeng,J. (2016) A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Res.*, **44**, e32.
56. Magnus,M., Antczak,M., Zok,T., Wiedemann,J., Lukasiak,P., Cao,Y., Bujnicki,J.M., Westhof,E., Szachniuk,M. and Miao,Z. (2020) RNA-puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Res.*, **48**, 576–588.
57. Yap,C.W. (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.*, **32**, 1466–1474.
58. Guo,J.C., Fang,S.S., Wu,Y., Zhang,J.H., Chen,Y., Liu,J., Wu,B., Wu,J.R., Li,E.M., Xu,L.Y., *et al.* (2019) CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. *Nucleic Acids Res.*, **47**, W516–W522.
59. Liu,Q., Chen,J., Wang,Y., Li,S., Jia,C., Song,J. and Li,F. (2021) DeepTorrent: a deep learning-based approach for predicting DNA N4-methylcytosine sites. *Brief Bioinform*, **22**, bbaa124.
60. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D., *et al.* (2016) Reference sequence (RefSeq) database

at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
61. Yates,A.D., Achuthan,P., Akanni,W., Allen,J., Alvarez-Jarreta,J., Amode,M.R., Armean,I.M., Azov,A.G., Bennett,R., Bhai,J., *et al.* (2020) Ensembl 2020. *Nucleic Acids Res.*, **48**, D682–D688.
62. Peng,H., Long,F. and Ding,C. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.*, **27**, 1226–1238.
63. Pan,X., Rijnbeek,P., Yan,J. and Shen,H.B. (2018) Prediction of RNA–protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *Bmc Genomics [Electronic Resource]*, **19**, 511.
64. Cheng,Z., Huang,K., Wang,Y., Liu,H., Guan,J. and Zhou,S. (2017) Selecting high-quality negative samples for effectively predicting protein-RNA interactions. *BMC Syst. Biol.*, **11**, 9.
65. Ma,X., Guo,J., Xiao,K. and Sun,X. (2015) PRBP: prediction of RNA-binding proteins using a random forest algorithm combined with an RNA-binding residue predictor. *IEEE/ACM Trans. Comput. Biol. Bioinform*, **12**, 1385–1393.
66. Wen,Z.Y., Liu,H.F., Shi,J.S., Li,Q.B., He,B.S. and Chen,J. (2020) ThunderGBM: fast GBDTs and random forests on GPUs. *J. Mach. Learn Res.*, **21**, 4389–4393.
67. Altmann,A., Tolosi,L., Sander,O. and Lengauer,T. (2010) Permutation importance: a corrected feature importance measure. *Bioinformatics*, **26**, 1340–1347.
68. Shen,W.X., Zeng,X., Zhu,F., Wang,Y.L., Qin,C., Tan,Y., Jiang,Y.Y. and Chen,Y.Z. (2021) Out-of-the-box deep learning prediction of pharmaceutical properties by broadly learned knowledge-based molecular representations. *Nat. Mach. Intell.*, **3**, 334–343.
69. MacPhillamy,C., Alinejad-Rokny,H., Pitchford,W.S. and Low,W.Y. (2022) Cross-species enhancer prediction using machine learning. *Genomics*, **114**, 110454.
70. Suresh,V., Liu,L., Adjeroh,D. and Zhou,X. (2015) RPI-pred: predicting ncRNA–protein interaction using sequence and structural information. *Nucleic Acids Res.*, **43**, 1370–1379.
71. Kim,C., Lee,H., Jeong,J., Jung,K. and Han,B. (2022) MarcoPolo: a method to discover differentially expressed genes in single-cell RNA-seq data without depending on prior clustering. *Nucleic Acids Res.*, **50**, e71.
72. Czerniak,T. and Saenz,J.P. (2022) Lipid membranes modulate the activity of RNA through sequence-dependent interactions. *Proc. Natl. Acad. Sci. U.S.A.*, **119**, e2119235119.
73. Riel,A.M.S., Rowe,R.K., Ho,E.N., Carlsson,A.C., Rappe,A.K., Berryman,O.B. and Ho,P.S. (2019) Hydrogen bond enhanced halogen bonds: a synergistic interaction in chemistry and biochemistry. *Acc. Chem. Res.*, **52**, 2870–2880.
74. Li,A., Zhang,J. and Zhou,Z. (2014) PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinf.*, **15**, 311.
75. Wang,L., Park,H.J., Dasari,S., Wang,S., Kocher,J.P. and Li,W. (2013) CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res.*, **41**, e74.
76. Kang,Y.J., Yang,D.C., Kong,L., Hou,M., Meng,Y.Q., Wei,L. and Gao,G. (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.*, **45**, W12–W16.
77. Chen,X.G., Liu,S. and Zhang,W. (2022) Predicting coding potential of RNA sequences by solving local data imbalance. *IEEE/ACM Trans. Comput. Biol. Bioinform*, **19**, 1075–1083.
78. Ramos,T.A.R., Galindo,N.R.O., Arias-Carrasco,R., da Silva,C.F., Maracaja-Coutinho,V. and do Rego,T.G. (2021) RNAmining: a machine learning stand-alone and web server tool for RNA coding potential prediction. *F1000Res.*, **10**, 323.
79. Zhang,Y., Jia,C., Fullwood,M.J. and Kwoh,C.K. (2021) DeepCPP: a deep neural network based on nucleotide bias information and minimum distribution similarity feature selection for RNA coding potential prediction. *Brief. Bioinform.*, **22**, 2073–2084.
80. Wucher,V., Legeai,F., Hedan,B., Rizk,G., Lagoutte,L., Leeb,T., Jagannathan,V., Cadieu,E., David,A., Lohi,H., *et al.* (2017) FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res.*, **45**, e57.
81. Hill,S.T., Kuintzle,R., Teegarden,A., Merrill,E., Danaee,P. and Hendrix,D.A. (2018) A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. *Nucleic Acids Res.*, **46**, 8105–8113.
82. Camargo,A.P., Sourkov,V., Pereira,G.A.G. and Carazzolle,M.F. (2020) RNAsamba: neural network-based assessment of the protein-coding potential of RNA sequences. *NAR Genom. Bioinform.*, **2**, lqz024.
83. Yang,S., Wang,Y., Zhang,S., Hu,X., Ma,Q. and Tian,Y. (2020) NCResNet: noncoding ribonucleic acid prediction based on a deep resident network of ribonucleic acid sequences. *Front. Genet.*, **11**, 90.
84. Yuan,J., Wu,W., Xie,C., Zhao,G., Zhao,Y. and Chen,R. (2014) NPInter v2.0: an updated database of ncRNA interactions. *Nucleic Acids Res.*, **42**, D104–D108.
85. Muppirala,U.K., Honavar,V.G. and Dobbs,D. (2011) Predicting RNA–protein interactions using only sequence information. *BMC Bioinf.*, **12**, 489.
86. Pan,X., Fan,Y.X., Yan,J. and Shen,H.B. (2016) IPMiner: hidden ncRNA–protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. *Bmc Genomics [Electronic Resource]*, **17**, 582.

Received: May 10, 2022. **Revised:** August 1, 2023. **Editorial Decision:** September 26, 2023. **Accepted:** October 10, 2023